International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Fake Image Detection Using Machine Learning

Mr. Anil Kumar K N, Asst. professor, Dept. of CSE, Rajeev Institute of Technology, Hassan, Karnataka Ms. Sinchana K S, Dept. of CSE, Rajeev Institute of Technology, Hassan, Karnataka
Ms. Sparsha D Y, Dept. of CSE, Rajeev Institute of Technology, Hassan, Karnataka Ms. Yashaswini B N, Dept. of CSE, Rajeev Institute of Technology, Hassan, Karnataka Ms. Sindhu H P, Dept. of CSE, Rajeev Institute of Technology, Hassan, Karnataka

Abstract — Design and implement a deepfake detection system capable of distinguishing authentic images from deepfake images that involve facial manipulation. This system should identify manipulated faces, thereby mitigating the harmful effects of deepfake technology. With the rapid advancement of image editing tools and generative technologies like deepfakes, the spread of manipulated or fake images has become a serious concern in areas ranging from social media to national security. Traditional methods of image verification are often inadequate due to the sophistication of modern forgeries.

I. INTRODUCTION

In the world of ever growing Social media platforms, Deepfakes are considered as the major threat of the AI. There are many Scenarios where these realistic face swapped deepfakes are used to create political distress, fake terrorism events, revenge porn, blackmail peoples are easily envisioned. Deep fakes are digitally manipulated images that appear deceptively real, often created with the aid of artificial intelligence.

II. LITERATURE SURVEY

[1] Title: Fake Image Detection Using Machine Learning
Year Of Publication: 2022 Author: A. Sharma and R. Mehta
Fake image detection using machine learning marks a critical advancement in digital content verification, enabling the identification of manipulated or synthetic images with high accuracy. Leveraging techniques such as convolutional neural networks (CNNs), deep learning models can analyze pixel- level features and uncover inconsistencies introduced by image editing tools or generative adversarial networks (GANs). These models operate on large datasets of real and fake images, learning to distinguish subtle artifacts and patterns that are imperceptible to the human eye. This technology addresses growing concerns over misinformation and digital forgery by providing automated, scalable solutions for media verification. The integration of fake image detection tools into social media platforms, news outlets, and forensic applications is set to enhance digital trust and protect against visual deception.
[2] Title: Fake Image Detection: A Machine Learning Perspective
Year Of Publication: 2022 Author: R. Kumar and M. Srinivas
Fake image detection using machine learning has emerged as a pivotal technique for safeguarding the integrity of digital content. By leveraging advanced algorithms like convolutional neural networks (CNNs), machine learning systems can detect visual anomalies and pixel-level inconsistencies introduced during image manipulation or deep fake generation. This technology addresses the limitations of manual verification by providing automated, real-time detection with high accuracy. Its deployment across social media platforms, legal forensics, and journalistic media is vital to combat misinformation and uphold visual authenticity in digital communication.
[3] Title: Deep Learning-Based Fake Image Detection Year Of Publication: 2023
Author: P. Verma and S. Rao
Deep learning-based fake image detection represents a technological leap in identifying altered or computer- generated images. Utilizing models such as ResNet and EfficientNet, these systems analyze spatial and frequency domain features to uncover traces of tampering, compression artifacts, or GAN-generated distortions. Operating on diverse datasets like Face Forensics++ and Deep Fake Detection, the models achieve high precision even under varying lighting conditions and image resolutions. The integration of deep learning detection tools into digital ecosystems ensures secure content authentication, supporting industries such as cybersecurity, media verification, and e-governance.
[4] Title: Fake Image Detection Using Convolutional Neural Networks

Year Of Publication: 2022
Author: A. Patel and R. Deshmukh
Fake image detection using convolutional neural networks (CNNs) represents a significant advancement in digital content authentication, enabling automated identification of tampered or computer-generated images. CNNs analyze visual data at multiple layers, capturing subtle pixellevel inconsistencies and artifacts introduced during image manipulation. This technology addresses the growing threat of visual misinformation by offering scalable, real-time analysis and decision-making. The integration of CNN-based detection systems into media platforms, law enforcement, and cyber security applications is poised to enhance digital trust and safeguard content integrity.

[5] Title: Machine Learning Techniques for Detecting Manipulated Images
Year Of Publication: 2023 Author: M. Singh and T. Bansal
Machine learning techniques for detecting manipulated images have emerged as crucial tools in the fight against digital deception. Leveraging algorithms like Support Vector Machines (SVM), Random Forest, and Deep Learning architectures, these systems identify irregularities in image texture, lighting, and compression patterns. Trained on diverse datasets, they can distinguish between original and altered visuals with high precision. This technology addresses the shortcomings of manual inspection and traditional forensic tools by offering automated, datadriven solutions.

III. Objectives

- Identifying Fake or Manipulated Images: Detect anomalies in images created through editing, manipulation, or generative models .
- Enhancing Trust in Digital Content: Provide tools to verify the authenticity of visual content and combat misinformation in social media, journalism, and other fields.
- Automating the Detection Process: Implement systems that perform real-time analysis of images to identify potential forgeries without manual intervention.
- Educating Users and Building Awareness: Provide insights and tools to help users identify fake images, fostering digital literacy.
- To explore and implement deep learning techniques for effective image detection. To evaluate the performance of popular models like CNNs and region-based methods (e.g., Faster R-CNN, YOLO).
- To analyze the impact of dataset size, preprocessing, and hyperparameter tuning on detection accuracy.
- To develop a prototype system capable of real-time image detection with high precision.

Scope : This project focuses on using deep learning algorithms for object detection in 2D static images. It includes model training, evaluation, and testing using standard datasets such as COCO or PASCAL VOC. The project will not address IMAGE processing, 3D detection, or deployment on embedded systems. Only supervised learning methods will be considered, and model optimization will be limited to commonly used techniques.

Applications

- Autonomous Vehicles: Detecting pedestrians, traffic signs, and other vehicles.
- Healthcare: Identifying abnormalities in medical imaging such as X-rays or MRIs.
- Security and Surveillance: Monitoring environments and recognizing suspicious activities.
- Retail and E-commerce: Visual product search and inventory management.

Agriculture: Detecting plant diseases or counting livestock via drone imagery.

IV. System Analysis

1. Problem Definition

With the rapid growth in image data across various sectors (security, healthcare, autonomous driving, etc.), traditional image processing methods are insufficient for detecting and classifying objects with high accuracy. There is a need for a system that can automatically detect objects within images using machine learning techniques.

2. Objectives of the System

- To develop a system that can automatically detect and recognize objects in images.
- To improve detection accuracy using deep learning models like CNN
- To process large datasets efficiently for real-time or near- real-time applications.
- To provide an interface for users to upload images and view detection results.

3. Scope of the System

- Input: Digital images (JPEG, PNG, etc.)
- Processing: Preprocessing, feature extraction, model training, and object detection
- Output: Labeled image with detected objects and their classes
- Application areas: Surveillance, medical diagnostics, traffic monitoring, etc.

4. Assumptions

- A labeled dataset for training (e.g., COCO, PASCAL VOC) is available.
- Sufficient computational resources (GPU-enabled system) are accessible.
- The user has access to the internet if cloud services are used for training/inference.

5. Feasibility Study

a. Technical Feasibility

Open-source libraries like TensorFlow, Keras, PyTorch, and OpenCV are available.
Pre-trained models (YOLO, Faster R-CNN, SSD) can reduce development time.

b. Operational Feasibility

The system can be used by researchers, engineers, and analysts with minimal training.
Web or desktop interfaces can simplify the interaction.

c. Economic Feasibility

Low cost if using open-source tool and free cloud- tier services.
Scalability available with cloud deployment options (AWS, Google Cloud, etc.).

6. Proposed System

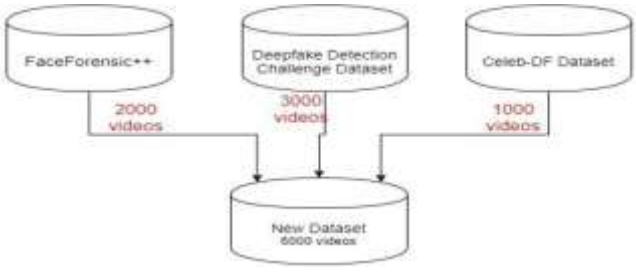
Architecture: Input Layer → CNN Backbone → Region Proposal Network (if applicable) → Detection Head → Output
Technologies: Python, TensorFlow/Py Torch, OpenCV, Flask (for web deployment)
Model Choices: YOLOv8, Efficient Det, or Faster R-CNN depending on accuracy- speed tradeoff .

Data-set Gathering

For making the model efficient for real time prediction. We have gathered the data from different available data-sets like FaceForensic++(FF)[1], Deepfake detection challenge(DFDC)[2], and Celeb- DF[3]. Futher we have mixed the dataset the collected datasets and created our own new dataset, to acurate and real time detection on different kind of images. To avoid the training bias of the model we have considered 50% Real and 50% fake images. Deep fake detection challenge (DFDC) dataset [3] consist of certain audio alerted image, as audio deepfake are out of scope for this paper. We preprocessed the DFDC dataset and removed the audio altered images from the dataset by running a python script. After preprocessing of the DFDC dataset, we have taken 1500 Real and 1500

Fake images from the DFDC dataset. 1000 Real and 1000 Fake images from the

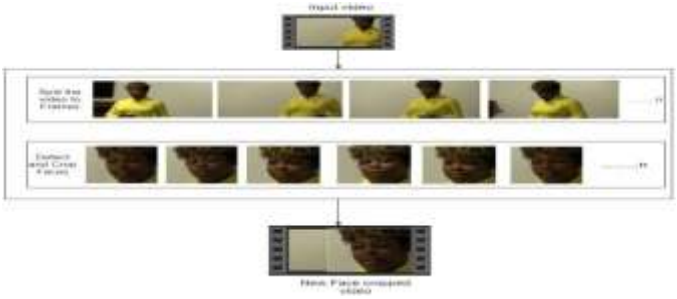
FaceForensic++(FF)[1] dataset and 500 Real and 500 Fake images from the CelebDF[3] dataset. Which makes our total dataset consisting 3000 Real, 3000 fake images and 6000 images in total. 2 depicts the distribution of the data-sets.



Pre-processing

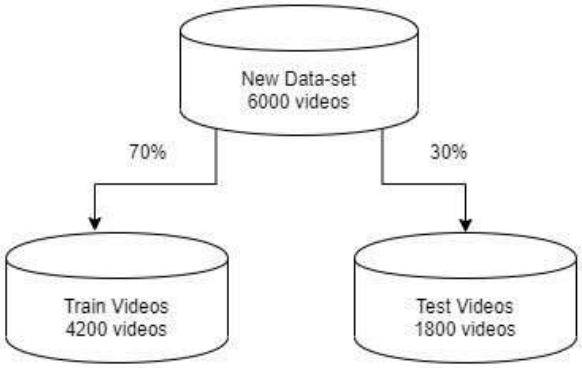
To maintain the uniformity of number of frames, we have selected a threshold value based on the mean of total frames count of each image. Another reason for selecting a threshold value is limited computation power. As a image of 10 second at 30 frames per second(fps) will have total 300 frames and it is computationally very difficult to process the 300 frames at a single time in the experimental environment. So, based on our Graphic Processing Unit (GPU) computational power in experimental environment we have selected 150 frames as the threshold value. While saving the frames to the new dataset we have only saved the first 150 frames of the image to the new image. To demonstrate the proper use of Long Short- Term
Memory (LSTM) we have considered the frames in the sequential manner

i.e. first 150 frames and not randomly. The newly created image is saved at frame rate of 30 fps and resolution of 112 x 112.



Data-set split

The dataset is split into train and test dataset with a ratio of 70% train images (4,200) and 30% (1,800) test images. The train and test split is a balanced split i.e 50% of the real and 50% of fake images in each split.



Model Architecture

Our model is a combination of CNN and RNN. We have used the Pre- trained ResNext CNN model to extract the features at frame level and based on the extracted features a LSTM network is trained to classify the image as deepfake or pristine. Using the Data Loader on training split of images the labels of the images are loaded and fitted into the model for training.

ResNext :

Instead of writing the code from scratch, we used the pre-trained model of ResNext for feature extraction. ResNext is Residual CNN network optimized for high performance on deeper neural networks. For the experimental purpose we have used resnext50_32x4d model. We have used a ResNext of 50 layers and 32 x 4 dimensions.

Following, we will be fine-tuning the network by adding extra required layers and selecting a proper learning rate to properly converge the gradient descent of the model. The

2048-dimensional feature vectors after the last pooling layers of ResNext is used as the

LSTM for Sequence Processing:

2048-dimensional feature vectors is fitted as the input to the LSTM. We are using 1 LSTM layer with 2048 latent dimensions and 2048 hidden layers along with 0.4 chance of dropout, which is capable to do achieve our objective. LSTM is used to process the frames in a sequential manner so that the temporal analysis of the image can be made, by comparing the frame at ‘t’ second with the frame of ‘t-n’ seconds. Where n can be any number of frames before t.

The model also consists of Leaky Relu activation function. A linear layer of 2048 input features and 2 output features are used to make the model capable of learning the average rate of correlation between eh input and output. An adaptive average polling layerwith the output parameter 1 is used in the model. Which gives the the target output size of the image of the form H x W. For sequential processing of the frames a Sequential Layer is used. The batch size of 4 is used to perform the batch training. A SoftMax layer is used to get the confidence of the model during predication.

V. System Architecture

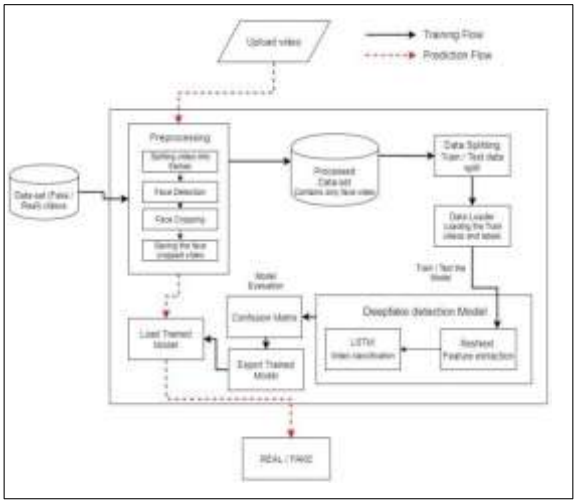


Fig 6.1: System Architecture

In this system, we have trained our PyTorch deepfake detection model on equal number of real and fake images in order to avoid the bias in the model. The system architecture of the model is showed in the Fig. In the development phase, we have taken a dataset, preprocessed the dataset and created a new processed dataset which only includes the face cropped images.

Creating deepfake images

To detect the deepfake images it is very important to understand the creation process of the deepfake. Majority of the tools including the GAN and autoencoders takes a source image and target image as input. These tools split the image into frames , detect the face in the image and replace the source face with target face on each frame. Then the replaced frames are then combined using different pre-trained models. These models also enhance the quality of image my removing the left-over traces by the deepfake creation model. Which result in creation of a deepfake looks realistic in nature. We have also used the same approach to detect the deepfakes. Deepfakes created using the pretrained neural networks models are very realistic that it is almost impossible to spot the difference by the naked eyes. But in
reality, the deepfakes creation tools leaves some of the traces

or artifacts in the image which may not be noticeable by the naked eyes. The motive of this paper to identify these unnoticeable traces and distinguishable artifacts of these images and classified it as deepfake or real image.

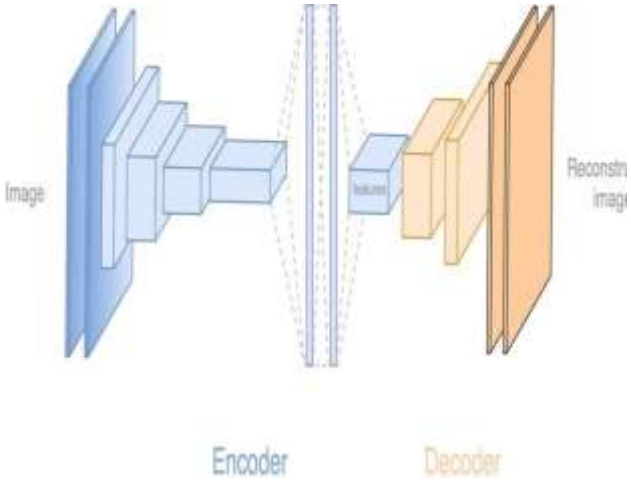
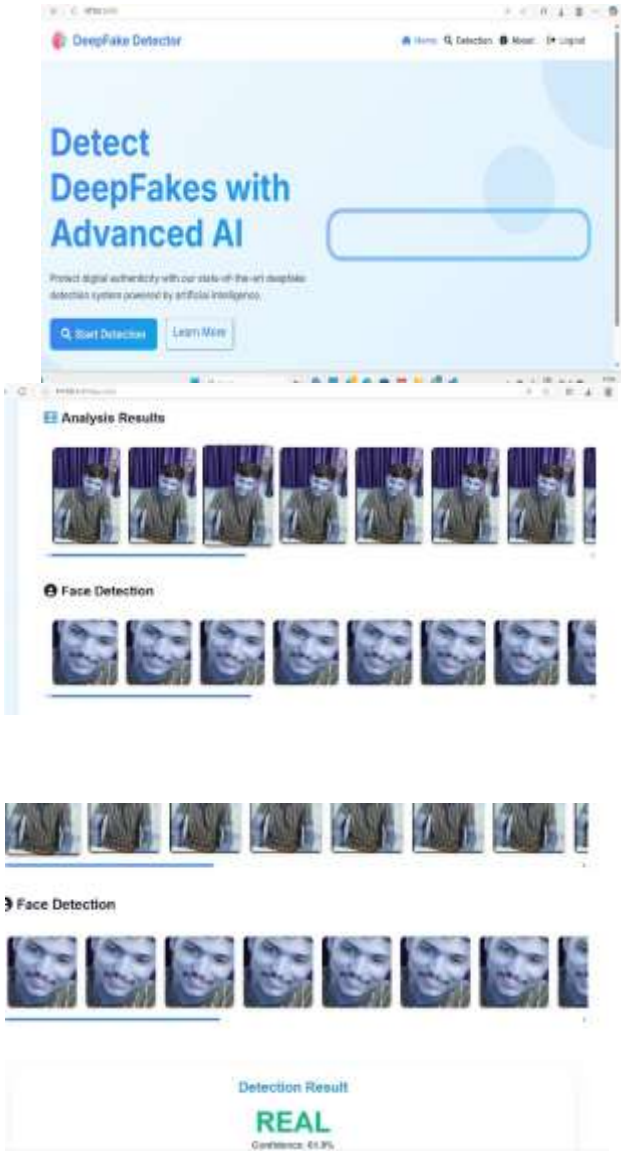
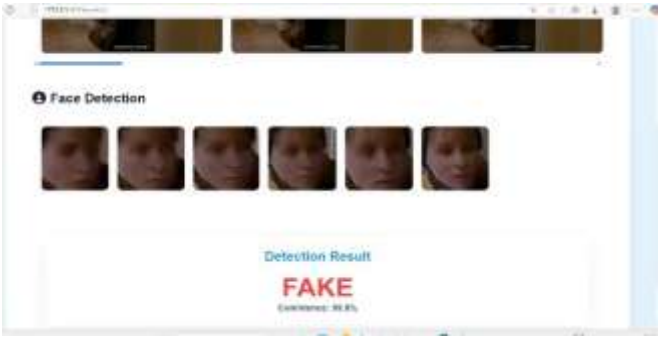


Fig :Deepfake generation

Face Swapped deepfake generation

VI. Result





CONCLUSION

DeepFake detection is a major need in today's world and needs considerable detection techniques as detecting deepfakes will become more challenging in the future. As deepfakes can have major social and political impact improvements should be made continuously in its detection techniques. For improving the performance, further research can be done on detecting temporal discrepancies and then using this combined information with features extracted from image processing module.

We presented a neural network-based approach to classify the image as deep fake or real, along with the confidence of proposed model. Our method is capable of predicting the

output by processing 1 second of image (10 frames per second) with a good accuracy. We implemented the model by using pre-trained Res Next CNN model to extract the frame level features and LSTM for temporal sequence processing to spot the changes between the t and t-1 frame. Our model can process the image in the frame sequence of 10,20,40,60,80,100.

REFERENCES

[1] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus This, Matthias Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images" in arXiv:1901.08971.

[2] Deepfake detection challenge dataset:https://www.kaggle.com/c/deepfake detection challenge /data Accessed on 26 March, 2020

[3] Yuezun Li , Xin Yang , Pu Sun , Honggang Qi and Siwei Lyu "Celeb-DF: A Large-scale

[4] Challenging Dataset for DeepFake Forensics" in arXiv:1909.12962 [4] Deepfake Image of Mark Zuckerberg Goes Viral on Eve of House A.I. Hearing : https://fortune.com/2019/06/12/deepfake- mark-zuckerberg/ Accessed on 26 March,2020

[5] 10 deepfake examples the terrified and amusedtheinternet:https://www.creativebloq.com/fea tures/deepfake-examples Accessed on 26 March,2020

[6] TensorFlow: https://www.tensorflow.org/ (Accessed on 26 March, 2020)